融合结构与文本特征的知识图谱关系预测方法研究*

■ 林泽斐[1,2]　欧石燕[1]

[1] 南京大学信息管理学院　南京 210093　2 福建师范大学社会发展学院　福州 350007

摘　要：[目的/意义]提出一种融合内部结构特征和外部文本特征的知识图谱关系预测新方法，旨在预测知识图谱中两实体间缺失关系的类型。[方法/过程]将关系路径和反映实体间关系的文本矩阵化，通过卷积神经网络学习与特定关系类型相关的结构和文本模式特征，在此基础上训练模型实现关系预测。[结果/结论]实验结果显示，该方法在评测数据集上的性能表现超过对照方法的水平，可有效提升知识图谱关系预测的性能。通过实际应用发现，该方法在知识服务中具有良好的应用价值。

关键词：知识图谱　关系预测　特征融合　深度学习

分类号：G254.29　TP391.1

DOI：10.13266/j.issn.0252-3116.2020.21.013

在大数据环境下，网络信息资源呈爆发式增长。网络信息资源中蕴含着大量有价值的知识，通过对这些知识的抽取和组织可以形成大规模知识库，进而为用户提供智能、高效的知识服务。知识图谱技术在这一背景下应运而生。知识图谱（knowledge graph，KG）旨在以结构化形式描述客观世界中的实体与实体间关系[1]，其将知识表示为形如 <h，r，t> 的三元组形式，其中 h，t，r 分别代表头部实体、尾部实体以及两者间关系。若将三元组中的实体表示为节点，将实体间的关系表示为边，可构建出一个图结构的知识网络。目前，知识图谱已在语义搜索、知识问答、智能推荐等领域得到了广泛应用[2]，显示出广阔的应用前景。近年来，出现了以 DBpedia、Freebase、CN-DBpedia 为代表的大规模知识图谱，尽管这些知识图谱中包含大量的三元组，但其知识仍远没有达到完备的程度，如 DBpedia 和 Freebase 中分别有 66% 和 71% 的人物实体缺少出生地信息[3]。在此背景下，知识图谱补全研究具有非常重要的意义。

知识图谱补全（knowledge graph completion，KGC）任务旨在减少知识图谱中的知识缺失，提高知识图谱中知识的完备程度[4]，譬如，图 1 是基于 CN-DBpedia 知识图谱中三元组数据[5]所构建的局部实体关系视图，利用该知识图谱中已有的显性知识可以推断出“马化腾”与“刘炽平”很可能存在同事关系，“深圳”有较大概率为“马化腾”与“刘炽平”的工作地，而这些隐性知识并未在该知识图谱中得以体现。因此，若能将知识图谱中的这些隐性知识补全，对于提升搜索引擎的语义搜索精度、完善知识问答和智能推荐系统的服务质量具有重要的价值。

具体而言，知识图谱补全又包括实体预测（entity prediction）和关系预测（relation prediction）两个典型的子任务[1]。其中，知识图谱实体预测任务旨在根据三元组的头实体和关系来预测并补全尾实体，或者根据尾实体与关系来预测补全头实体，即 <h，r，?> 和 <?，r，t>；知识图谱关系预测任务则是根据给定三元组中的头实体和尾实体，预测出两实体间缺失关系的类型，即 <h，?，t>[6]。本研究主要围绕知识图谱中的关系预测问题展开。

知识图谱的内部结构以及知识图谱的外部文本均可作为关系预测的依据，例如，根据图 1 中“马化腾”到“刘炽平”间的路径结构，可以推断“马化腾”和“刘炽平”间可能存在着“同事”关系，若有文本提及“刘炽平协助马化腾管理公司的日常运营”，也可以做出类似的推理，而上述两类信息相结合可以帮助计算机更加准确地识别出实体间关系。因此，内部结构和外部文本在知识图谱关系预测任务中具有互补性，但当前将二

* 本文系国家社会科学基金重点项目"基于关联数据的学术文献内容语义发布及其应用研究"（项目编号：17ATQ001）研究成果之一。
作者简介：林泽斐（ORCID：0000-0001-8637-7359），讲师，博士研究生；欧石燕（ORCID：0000-0001-8617-6987），教授，博士生导师，通讯作者，E-mail：oushiyan@nju.edu.cn。
收稿日期：2020-06-09　修回日期：2020-08-19　本文起止页码：99-110　本文责任编辑：易飞

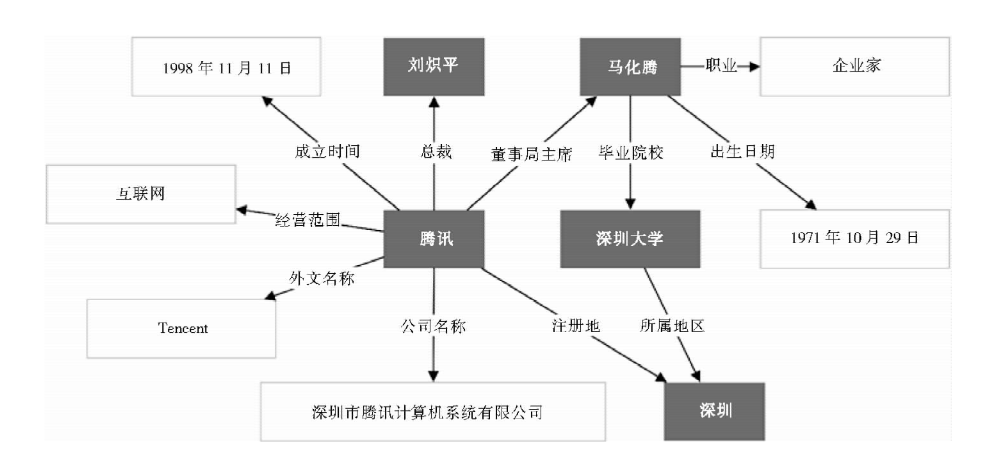


图 1　CN-DBpedia 知识图谱中的局部实体 – 关系视图

者结合进行关系预测的研究仍十分缺乏。针对这一问题,本研究将探索一种新的知识图谱关系预测方法,其创新之处在于设计了一种对实体间关系路径结构特征进行学习的深度学习模型,并将外部文本的特征集成到该模型中,从而构建出一个融合知识图谱内部结构与外部文本特征的关系预测集成模型。

1　相关工作

当前,针对知识图谱中缺失关系的预测有两种典型方法:基于知识表示学习的方法和基于关系路径的方法。

知识表示学习是利用机器学习技术学习出知识图谱中实体和关系的低维稠密向量表示[7]。以 TransE 模型为代表的翻译模型是当前知识表示学习的主流模型[8]。TransE 模型[9]将关系向量 r 视为头实体向量 h 和尾实体向量 t 之间的平移向量,如 h(腾讯) + r(董事会主席) ≈ t(马化腾),通过最小化 h + r 与 t 之间的距离训练出实体和关系的向量表示。得到每一实体和关系的向量表示后,通过寻找使 h + r 与 t 之间的距离尽可能小的 r,即可预测出头尾实体间的关系类型。TransE 在处理"一对多""多对多"等复杂关系时存在局限性,故一些学者提出了 TransE 模型的变体形式,如 TransH、TransR 和 TransD 等,其中 TransH[10]为不同关系指定不同的超平面,从而使一个实体在不同的关系下拥有不同的向量表示,而 TransR[11]和 TransD[12]模型则通过在不同语义空间内表示实体和关系来解决上述问题。除翻译模型外,语义匹配模型也是知识表示学习中的一类经典模型,此类模型将实体和关系映射到向量空间中,通过相似度评价函数来学习实体和关系的隐含特征[13],代表性工作包括 RESCAL 模型[14]及其改进模型 DistMult[15]和 ComplEx[16]等。基于知识表示学习的关系预测方法其优势在于预测准确率较高,且具有较强的通用性,训练出的实体和关系向量不仅适用于关系预测任务,也适用于实体预测和三元组分类等其他任务;不足之处在于其主要关注以三元组为单元的局部信息,而较少关注知识图谱的网状结构与关系间的逻辑推理。

关系路径是知识图谱中两实体间的连通路径,这些连通路径也能反映出实体之间的语义关系。基于关系路径的方法使用关系路径作为特征来对实体间的关系类型进行预测。卡内基梅隆大学的 N. Lao 等所提出的路径排序算法(path ranking algorithm, PRA)[17]是此类方法的典型代表。该算法从知识图谱中提取出所有带有关系 r 的三元组集合,以集合中每个三元组的头实体作为起点进行随机游走,若三步内可到达尾实体,则将该游走路径 π 作为关系 r 的一条关系路径用于模型训练,通过逻辑回归算法在训练集上学习出各路径相对于关系 r 的权重。接下来,通过一种路径约束的随机游走(path-constraint random walk)方法来计算头实体沿着路径 π 到达尾实体的转移概率。最后,使用一个综合路径权重和转移概率的评分函数来评估实体对间存在关系 r 的可能性。基于关系路径的方法利用了知识图谱的网状结构与实体关系间的逻辑推理,具有良好的可解释性,但 PRA 对于关系路径的学习能力仍较为有限,其关系预测性能相比 TransE 等主流模型有一定差距[18]。

上述两种方法均基于知识图谱内部结构特征进行关系预测,区别在于前者关注以三元组为单元的微观结构,而后者关注以路径为单元的宏观结构。除结构特征外,知识图谱外部的文本信息也可作为知识图谱关系预测的依据。近年来,一些研究者开始利用外部文本信息改善知识图谱补全的效果,相关研究多是将

实体描述文本（主要出自百科词条摘要）与知识表示学习方法相结合，如J. Xu等将实体描述文本的向量表示与TransE相结合，实现联合表示学习[19]；T. Long等利用词条描述文本的平均词向量来初始化实体向量，以改善TransE模型的性能[20]。但是，就知识图谱关系预测任务而言，同时包含两实体指称的句子显然较实体描述文本更能直接地反映出两实体间的关系，而当前相关研究中并未使用到此类文本。

此外，深度学习是近年来机器学习领域的突破点，特别是基于卷积神经网络（convolutional neural network, CNN）的深度学习模型由于具有优秀的特征表示能力，已被广泛应用于计算机视觉等领域，但在知识图谱关系预测任务中却较少被应用。近年来，有学者开始将CNN应用到知识图谱补全任务中。T. Dettmers等于2018年提出的ConvE模型[21]是第一个将CNN应用于知识图谱补全的模型。ConvE模型将一维的头实体向量和关系向量分别变形为2D矩阵，并通过纵向拼接形成新的矩阵作为卷积层的输入，然后使用3 × 3大小的卷积核对输入矩阵进行卷积操作，以此进行模型训练。ConvE在知识图谱补全中具有良好的预测性能，但该模型并未考虑知识图谱中的关系路径信息和文本信息。通过检索，目前尚未发现利用深度学习模型融合路径结构特征与文本特征实现知识图谱关系预测的相关研究。

综上所述，当前的知识图谱关系预测主要基于知识图谱的结构特征来进行。一些研究开始将结构特征与文本特征相集成以提升预测性能，但此类研究并未关注以关系路径为代表的宏观结构特征，所采用的文本的类型也存在一定的局限性；同时，深度学习技术在该领域的应用仍十分有限。鉴于多特征融合和深度学习技术在知识图谱关系预测领域具有较大的潜力，本研究拟对该问题进行探索，提出一种基于卷积神经网络并融合知识图谱内部结构与外部文本特征的知识图谱关系预测新方法。

2 知识图谱关系预测模型的设计

本研究提出了一个基于卷积神经网络，融合知识图谱内部结构特征和外部文本特征的知识图谱关系预测模型ConvF。该模型将知识图谱关系预测视为一个多标签分类问题，模型的输入为反映两实体间关系路径的结构特征和反映两实体间关系的文本特征，模型的输出为两实体间存在不同关系类型（在指定的关系类型集合中）的概率，其功能结构如图2所示：

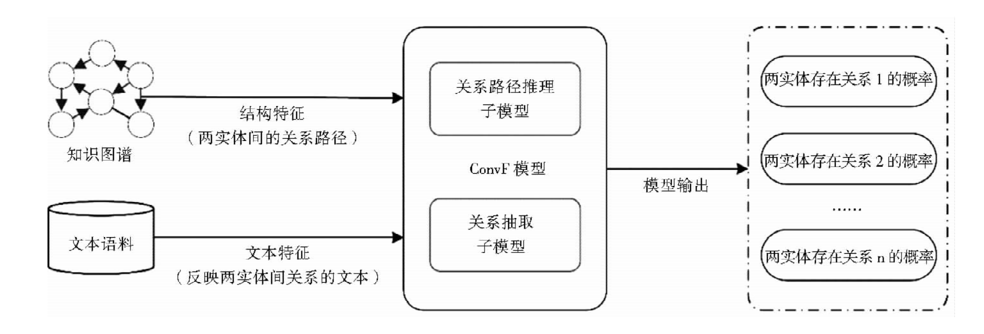


图2 ConvF模型的功能结构

ConvF模型由关系路径推理和关系抽取两个子模型构成，其中，关系路径推理子模型用于学习知识图谱内部的关系路径结构特征；关系抽取子模型则用于学习与实体对相关的外部文本的特征。在本节中将分别介绍这两个子模型的设计以及它们如何融合成为ConvF关系预测模型。

2.1 关系路径推理子模型

关系路径推理子模型通过学习知识图谱内部的关系路径特征进行关系预测。该子模型的设计思路为：知识图谱中存在特定直接关系的两实体间往往还存在着多条连通路径（即关系路径），可将这些关系路径作为该直接关系的模式特征，通过机器学习算法对这些模式特征进行学习，若待预测的两实体间存在类似的关系路径结构，即可判断出两实体间有较大概率存在此类直接关系。如图3所示，用于模型训练的实体 e_t 相对于 e_h 的直接关系为Friend，且从 e_h 到 e_t 存在两条关系路径（Enemy，Enemy）和（Friend，Friend），则两条路径可联合作为Friend关系的模式特征；待预测关系的两实体 $e_{h'}$ 和 $e_{t'}$ 间的路径结构稍有不同，但同样存在类似的路径结构模式，故可推测 $e_{t'}$ 有较大概率为 $e_{h'}$ 的Friend。

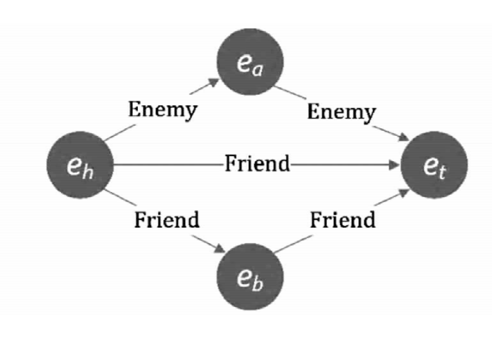


图 3 – a 训练实体对间的关系路径结构 图 3 – b 待预测关系的实体对间的关系路径结构

接下来对上述问题进行形式化表述。知识图谱 G = {E, R} 可被视为一个有向多重图(Directed Multigraph),其中,E 为知识图谱中的实体集合;R 为知识图谱中的关系集合;$< e_h, e_t >$ ($e_h, e_t \in$ E) 为知识图谱中的一个实体对;$r_{h\to t}$ ($r_{h\to t} \in$ R) 为 e_t 相对于 e_h 的关系。若将 e_h 视为路径的起点,e_t 视为路径的终点,$\pi_{h\to t} = (e_h, r_1, e_1, r_2, e_2, \cdots, r_n, e_t)$ 代表从实体 e_h 到实体 e_t 间的一条关系路径,该路径可简写为关系序列的形式,即 $\pi_{h\to t} = (r_1, r_2, \cdots, r_n)$,其中 n 称为关系路径的跳数(hop);$e_h$ 到 e_t 间的所有 n 跳关系路径所构成的集合记为 $\Pi^n_{h\to t}$。

由于知识图谱为有向图,故路径中的关系 r 存在方向之分。本研究将路径中指向 e_t 方向的关系定义为正向关系(记为 r),指向 e_h 方向的关系定义为反向关系(记为 r^{-1})。特别地,$\pi^d_{h\to t} = (e_h, r_{h\to t}, e_t)$ 称为 e_h 到 e_t 的正向直接关系路径(上标 d 表示直接关系),该路径所包含的唯一关系就是待学习或预测的目标关系。

关系路径推理子模型使用卷积神经网络(CNN)来学习关系路径的模式特征。由于 CNN 模型的输入为 2D 矩阵,故首先需要将关系路径转换为矩阵形式。矩阵转换采用如下方法:将知识图谱 G 中的每一个关系 r 表示为一个 k 维向量,将一条 n 跳关系路径中各关系的 k 维向量进行纵向拼接,即可得到一个 n × k 大小的矩阵。从 e_h 到 e_t 可能存在多条的 n 跳关系路径,这些路径的集合可通过深度优先搜索(DFS)或广度优先搜索(BFS)算法遍历生成。n 跳关系路径的集合同样可以表示为矩阵形式,这一矩阵由多个 n 跳关系路径矩阵纵向拼接得到,相关公式可表示为:

$$PM_{n,i} = \overrightarrow{r_{i,1}} \oplus^{\downarrow} \overrightarrow{r_{i,2}} \oplus^{\downarrow}, \cdots, \oplus^{\downarrow} \overrightarrow{r_{i,n}} \quad \text{公式(1)}$$

$$M_n = PM_{n,1} \oplus^{\downarrow} PM_{n,2} \oplus^{\downarrow}, \cdots, \oplus^{\downarrow} PM_{n,\theta_n} \quad \text{公式(2)}$$

其中,$PM_{n,i}$ 为实体 e_h 到 e_t 间第 i 条 n 跳关系路径 $\pi_{n,i}$ ($\pi_{n,i} \neq \pi^d_{h\to t}$) 的矩阵表示;$\overrightarrow{r_{i,1}}$, $\overrightarrow{r_{i,2}}$, $\cdots$, $\overrightarrow{r_{i,n}}$ 为该路径中关系的行向量表示;$\oplus^{\downarrow}$ 代表行向量或矩阵的纵向拼接;M_n 为实体 e_h 到 e_t 间 n 跳关系路径集合的矩阵表示;θ_n 为超参数,代表人工设定的矩阵中所包含的最大路径数量。得到不同跳数的关系路径集合矩阵 M_n (n = 1, 2, 3, ⋯)后,即可利用 CNN 模型从这些矩阵中学习出针对 $r_{h\to t}$ 关系的关系路径模式特征。

下面以一简化实例说明关系路径集合矩阵的构建。假设存在图 4 所示的知识图谱,e_h、e_t、e_a、e_b 为知识图谱中的实体,该图谱的关系集合表示为 R = {mother, son, brother, grandson},现需要对 $r_{h\to t}$ (即 son 关系)的模式特征进行学习。从 e_h 到 e_t 存在两条单跳关系路径(son)和(mother^{-1}),以及两条两跳关系路径(mother, grandson)和(mother^{-1}, brother)。其中(son)为 e_h 到 e_t 的正向直接关系路径 $\pi^d_{h\to t}$,该路径不被包含在构建的矩阵中。若 **mother**, **mother**$^{-1}$, **son**, **son**$^{-1}$, **brother**, **brother**$^{-1}$, **grandson**, **grandson**$^{-1}$分别为上述 4 种关系及其反向关系的行向量表示,可基于公式(1)和(2)构建出单跳关系路径矩阵 M_1 和双跳关系路径矩阵 M_2。M_1 和 M_2 可共同作为 son 关系的训练样本输入数据用于模型训练。需要说明的是:①本例以人物知识图谱为例说明关系路径集合矩阵的构建,但该方法同样适用于其他各类通用知识图谱中的关系预测;②在该例中从 e_h 到 e_t 仅存在单跳和两跳关系路径,在实际的知识图谱中,两实体间还可能存在 3 跳或 3 跳以上的关系路径,这些关系路径均可使用相似的方法构建关系路径集合矩阵;③由于我们将知识图谱视为一个有向多重图,即知识图谱中的实体间可以存在多重关系,如 e_h 可以既是 e_t 的母亲,又是 e_t 的老师,故单跳关系路径集合矩阵也可以包含超过一条关系路径。

构建关系路径集合矩阵的一个关键问题是关系向量如何表示。一种简单的策略是为每一种关系类型生成一个独热(one-hot)形式的向量,如将 son 关系表示为(0, 1, 0, 0),但若知识图谱中的关系类型数量较多,使用此方法生成的关系路径集合矩阵将极为稀疏,这大大提高了训练过程中的资源消耗;另一种解决策略

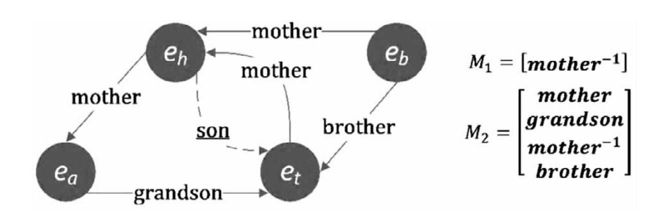


图 4　关系路径集合矩阵构建实例

是通过知识表示学习方法(如 TransE)为每一种关系类型生成一个预训练的低维稠密向量表示,使用该方法可大大降低资源消耗,且生成的向量具有一定的语义信息,即含义相似的关系具有空间距离接近的向量表示。本研究采用后一种方法生成关系向量。另外,关系的方向显然将影响到路径推理的结果,因此对于不同方向的关系,需要生成不同的向量表示,而知识表示学习方法所生成的向量并未对关系在路径中的方向进行区分。针对这一问题,我们采用了一种简单的解决策略,即对于 n 跳关系路径(n>1),在每一关系向量的前部拼接一个 one-hot 形式的方向向量,若关系方向为正向,则方向向量为(1,0);反之,则为(0,1),从而使生成的关系向量具有方向特征。对于单跳关系路径,由于我们不向矩阵中添加正向直接关系路径,故矩阵中只存在反向关系,无需添加方向向量。

关系路径推理子模型将关系预测视为一个多标签分类问题,其输入为实体 e_h 到实体 e_t 间的 n 跳关系路径集合矩阵(n=1,2,3,...),输出为 $r_{h \to t}$ 属于不同关系类别的概率。该模型由输入层、卷积层、池化层、全连接层和输出层 5 部分构成,如图 5 所示:

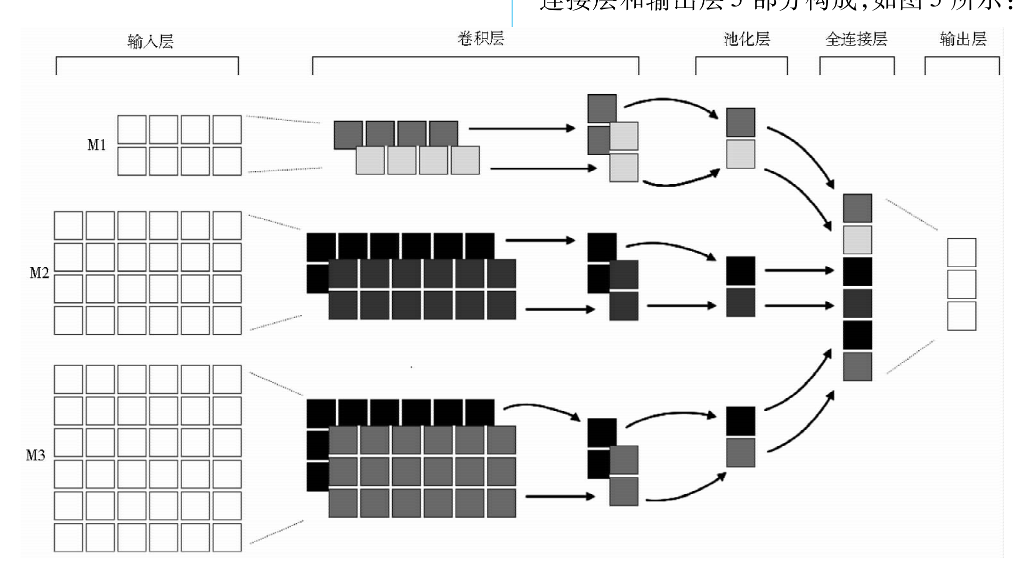


图 5　关系路径推理子模型结构示意

(1)输入层包含不同跳数的关系路径集合矩阵。其中,关系路径集合矩阵 M_n 的宽度 w_n 取决于所使用的关系向量的维度 k。由于单跳关系路径不需要添加方向向量,故 $w_1 = k$;对于其他跳数的关系路径集合矩阵,其宽度为 $k+2$。这些矩阵的高度为 $n \times \theta_n$,其中 n 为矩阵中路径的跳数,超参数 θ_n 为每个矩阵所允许包含的 n 跳路径的最大数量。若知识图谱中实体 e_h 到实体 e_t 实际存在的 n 跳路径数量 $>\theta_n$,则多余的路径将被随机剔除;若实际存在的 n 跳路径数量 $<\theta_n$,则多余的行以零向量进行填充。

(2)卷积层用于实现路径特征的提取。对于每个 n 跳关系路径集合矩阵,分别设置 F_n 个卷积核进行卷积运算,卷积核大小为 $n \times w_n$。卷积核的滑动步长被设置为 n,这样卷积核每次滑动得到的感受野(receptive field)均对应于一条完整关系路径的特征矩阵。卷积操作完成后,每个卷积核将生成一个一维的特征图。

(3)池化层对每个特征图进行全局最大池化(global max pooling),共得到 $\sum_{i=1}^{n} F_i$ 个标量值。

(4)全连接层将池化层生成的所有标量值拼接为一个向量,并以全连接方式与输出层相连。为减少训练过程中过拟合情况的发生,全连接层采用了 Dropout 策略[22]。

(5)输出层使用 sigmoid 激活函数计算实体对属于各关系类别的概率,并通过二元交叉熵(binary cross-entropy)函数计算模型损失。

2.2　关系抽取子模型

关系抽取子模型通过学习知识图谱外部的文本特征进行关系预测。关系抽取(relation extraction)是信

息抽取中的一项重要任务，其目标是给定一段包含两实体指称的文本，根据文本判断出两实体间的关系类型[23]。关系抽取子模型借鉴了关系抽取的思路：在知识图谱的外部文本（如百科词条）中，存在着大量包含两个或两个以上实体指称（mention）的句子，通过将三元组中的实体与这些语句中的实体指称对齐，寻找同时包含头实体和尾实体指称的句子集合，并将句中的实体指称使用掩码进行替换，即可从这些句子集合中学习出与特定关系类型有关的文本模式特征，例如，假设存在三元组 <鲁迅，作品，狂人日记>，而相关百科词条中提及"鲁迅发表白话小说《狂人日记》""《狂人日记》是鲁迅创作的白话小说"，分别将句中头实体和尾实体指称替换为掩码 <head> 和 <tail>，则" <head> 发表白话小说 <tail>"" <tail> 是 <head> 创作的白话小说"均可反映出"$r_{h\to t}$ = 作品"这一关系。

关系抽取子模型也将关系预测视为一个多标签分类问题，并同样采用了 CNN 作为分类器，其结构借鉴了 TextCNN 文本分类模型[24]的设计模式。图 6 为关系抽取子模型的结构示意图，该子模型同样由 5 部分构成：

（1）输入层为文本的矩阵表示，矩阵中的每行为一个单词的词向量表示。若文本由 n 个单词构成，每个单词的词向量维度为 k 维，则输入层为一个 n×k 的矩阵。

（2）卷积层设置不同高度的卷积核完成关系模式特征的提取。所有卷积核的宽度均为 k，各卷积核的高度为超参数。卷积核在文本矩阵中纵向滑动进行卷积操作，滑动步长为 1。

（3）池化层、全连接层和输出层的设计与关系路径推理子模型相同，在此不作赘述。

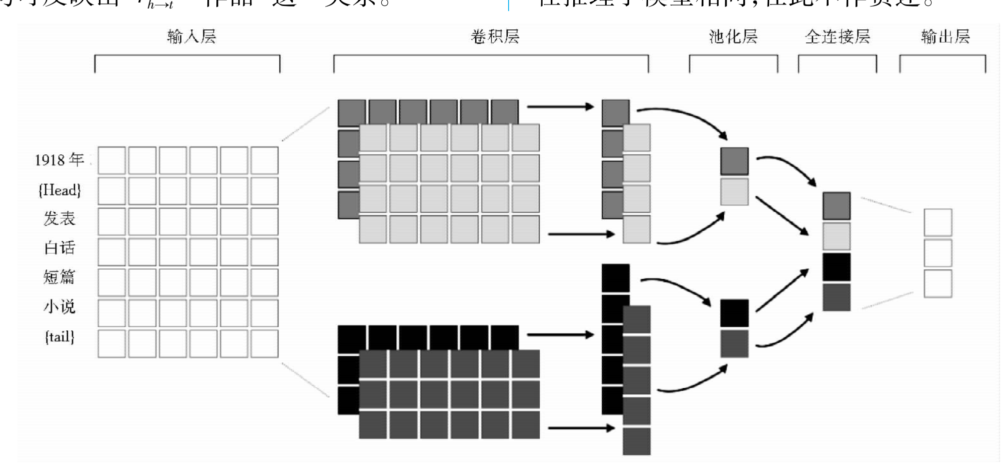


图 6 关系抽取子模型结构示意

2.3 融合模型

上述两个子模型可以分别独立地实现关系预测，但是在实际的关系预测任务中，仅凭关系路径或文本信息其中之一往往并不足以做出准确的预测，例如，若实体 B 为实体 A 的朋友，实体 C 为实体 B 的朋友，根据（朋友，朋友）这一关系路径，可以推断 C 可能为 A 的朋友，但这一推断并不确定，因为"A 朋友的朋友"未必一定为"A 的朋友"；但是，若有文本提及"C 在生活中对 A 予以照料"，综合上述路径和文本信息综合判断，则能够得出"C 为 A 的朋友"这一较为确定的结论。

为了融合关系路径推理子模型和关系抽取子模型实现综合推理，我们提出了一个融合模型（ConvF）实现二者的集成。图 7 为 ConvF 模型的结构示意图，该融合模型的前半部分由关系抽取和关系路径推理子模型的输入层、卷积层和池化层共同构成。在全连接层，我们将两个子模型的神经元进行拼接处理，使该层神经元既包含基于文本的特征表示，也包含基于关系路径的特征表示。增加隐藏层可以在一定程度上提高神经网络模型的特征表示能力[25]，故我们在拼接形成的全连接层和输出层中间添加了一个全连接隐藏层以获得更低的训练误差。

ConvF 模型的使用过程分为预处理、训练和预测 3 个阶段：①在预处理阶段，将已知关系的三元组 <h，r，t> 视为一个训练样本，对于每一个训练样本中的实体对 <h，t>，通过图搜索算法生成图中对应节点 e_h 到 e_t 间的关系路径集合 $\Pi_{h\to t}$（n =1，2，3，…），使用 2.1 节所述方法将关系路径集合转换为关系路径集合矩阵 M_n（n =1，2，3，…）；同时，从外部文本中寻找同时包含 h 和 t 指称的句子集合，使用 2.2 节所述方法生成文本的矩阵表示；②在训练阶段，将每一个训练样本 <h，r，t> 的关系路径集合矩阵和文本矩阵共同作为模型输

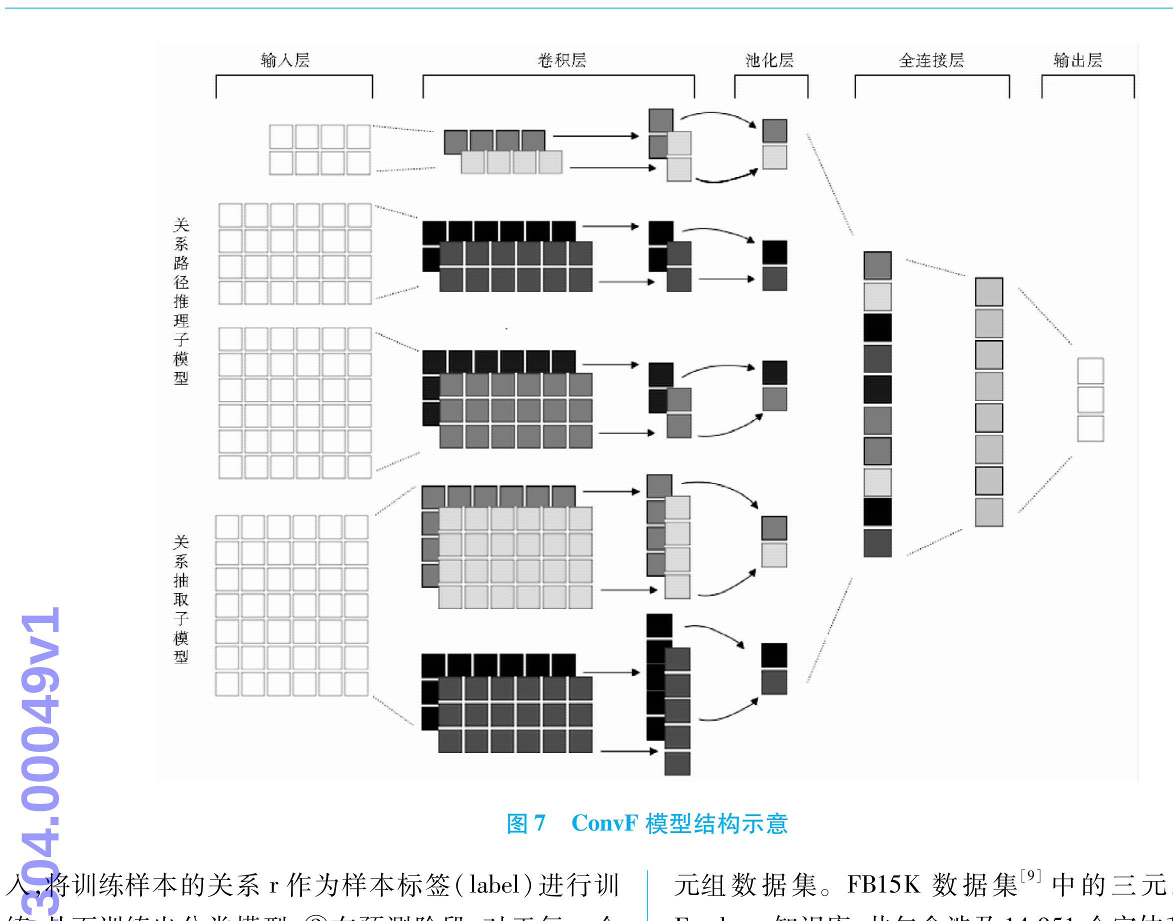


图7　ConvF模型结构示意

入，将训练样本的关系r作为样本标签(label)进行训练，从而训练出分类模型；③在预测阶段，对于每一个待预测关系的实体对＜h′,t′＞，使用与训练过程相同的方法生成针对h′和t′的关系路径集合矩阵和文本矩阵，将这些矩阵作为模型的输入数据，模型将输出h′,t′的关系属于各关系类型的不同概率。需要说明的是，在预测阶段，ConvF并不强制要求输入数据同时包含关系路径集合矩阵和文本矩阵。待预测关系的实体对可能存在信息缺失，即一些实体对可找到两实体间的关系路径，但无法匹配到同时包含两实体对的句子集合，而一些实体对则反之。对于此类存在信息缺失的实体对，可仅使用关系路径集合矩阵或文本矩阵其中之一作为模型输入，无法获取对应信息的矩阵则以零矩阵填充，模型可基于有限的信息输出预测结果。

ConvF相对于传统关系预测模型的主要优势在于：①模型可综合结构和文本信息进行关系预测；②该模型可通过深度学习实现更加有效的关系路径模式特征学习。在下一节中，将对模型的性能予以评测。

3　实验与结果分析

3.1　评测数据集

本研究使用FB15K和FB15K237数据集的衍生数据集FB15K－T和FB15K237－T进行模型性能的评测。FB15K和FB15K237是关系预测任务中常用的三元组数据集。FB15K数据集[9]中的三元组提取自Freebase知识库，共包含涉及14 951个实体和1 345种关系的50余万个三元组样本。K. Toutanova等于2015年指出FB15K测试集中的大量三元组可通过反向关系映射至训练集中的三元组，如训练集中包含＜猫，上位类，猫科动物＞，而测试集包含＜猫科动物，下位类，猫＞，故仅需要一个简单的基于规则的模型即可实现与主流模型相似的性能[26]。针对这一问题，K. Toutanova等引入了FB15K237数据集[27]。FB15K237从FB15K中去除了大多数存在直接反向关系的实体对，从而使评测任务更专注于非平凡的知识推理。

FB15K和FB15K237仅包含知识图谱的结构信息，为了测试结构与文本相结合的知识图谱关系预测方法，需要在FB15K和FB15K237数据集基础上添加与三元组相关的外部文本信息。鉴于Freebase知识库中的大部分知识均抽取自维基百科，故基于Freebase生成的FB15K和FB15K237数据集中的多数实体可与维基百科词条实现映射。因此，我们首先将FB15K和FB15K237数据集中的实体映射到对应的维基百科词条；然后，对于数据集中的每一个三元组，从头实体和尾实体词条页中寻找同时包含两实体指称的句子集合(将句中的实体指称使用掩码替换)；最后将句子集合作为各三元组的对应文本，生成形如＜头实体，尾实体，实体间关系，文本＞的四元组。经统计，在FB15K

和 FB15K237 中有约 1/3 的三元组样本可匹配到对应的文本信息，我们选择 FB15K 和 FB15K237 数据集中的此类样本生成相应的四元组样本，形成 FB15K－T 和 FB15K237－T 数据集。两个评测数据集及其来源数据集的基本情况如表 1 所示：

表 1　评测数据集及其来源数据集的基本情况

数据集(单位)	FB15K	FB15K－T	FB15K237	FB15K237－T
关系数(种)	1 345	1 177	237	226
实体数(个)	14 951	13 060	14 541	12 619
训练样本数(元组)	483 142	143 277	272 115	82 095
验证样本数(元祖)	50 000	14 915	17 535	4 436
测试样本数(元组)	59 071	17 708	20 466	5 200

3.2　评测指标

关系预测任务的评价通常采用基于排序的评价方法，即给定一个测试集中的实体对＜h，t＞，预测出带有偏序的关系列表 $l=[r_1,r_2,\cdots,r_n]$，若测试集标注的实体对真实关系标签 r^* 在 l 中排名靠前，则说明预测性能较好。我们参考本领域评测的常用作法，使用 MRR 和 Hits@n 作为评测指标。MRR(mean reciprocal rank)为正确结果在预测列表中排名的倒数的平均值；Hits@n 指正确结果在预测结果列表中前 n 名的比例，两者的计算公式分别为：

$$MRR=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \qquad 公式(3)$$

$$Hits@n=\frac{1}{N}\sum_{i=1}^{N}\sigma(rank_i\leqslant n) \qquad 公式(4)$$

其中，N 为待预测关系类型的实体对数量；$rank_i$ 为待预测关系类型的第 i 个实体对的正确关系类型在预测出的关系列表中的排名位置；σ 为指示函数，其定义为：

$$\sigma(rank_i\leqslant n)=\begin{cases}1 & 当\ rank_i\leqslant n\\ 0 & 当\ rank_i>n\end{cases} \qquad 公式(5)$$

MRR 和 Hits@n 的值域均位于 0 至 1 区间，其值越接近 1，代表模型具有越好的预测性能。由于实体对＜h，t＞在知识图谱中可能存在不止一种正确的关系类型，若 r^* 在 l 中排名为 k，$r_j(0<j<k)$ 实际上也可能为正确的结果，故本领域研究通常将 MRR 和 Hits@n 进一步细分为 raw 和 filter 两个类别，前者指不从预测结果列表中去除其他正确结果的原始评测指标，后者指从预测结果列表中去除其他正确结果(从而使预测结果的排名更加靠前)后得到的评测指标。本研究在评测时同样采用了这种分类方法。

3.3　模型训练

本节阐述模型的训练过程与训练参数设置。为了进行模型训练，首先需要利用训练集三元组构建一个有向多重图，以利用该图生成关系路径。本研究基于 Neo4j 图数据库实现图的构建，方法为：将数据集包含的实体转换为图中的节点，利用训练集中的三元组在 Neo4j 中创建边，以此构建出一个图。两实体间关系路径的生成可采用成熟的图搜索算法，如深度优先搜索(DFS)或广度优先搜索(BFS)算法。Neo4j 已实现了基于 DFS 的图搜索算法，故我们直接使用 Neo4j 生成两实体间的关系路径集合。需要说明的是，在两实体间可能存在多条关系序列相同的路径，例如实体 e_h 和 e_t 间可能存在关系路径 $\pi_1=(e_h,r_1,e_1,r_2,e_t)$ 与 $\pi_2=(e_h,r_1,e_2,r_2,e_t)$，两者的区别仅为途经的中间节点不同。在本任务中，仅关心路径中的关系 r_i，而 π_1 与 π_2 均可表示为关系序列(r_1,r_2)，故我们将其合并为一条路径加入集合；另外，遍历两节点间的所有关系路径具有较高的时间复杂度，为了提高效率我们将路径搜索的范围限制在 DFS 算法随机匹配到的前 m 条路径(m 为超参数)，对于 1－3 跳关系路径，将 m 分别设置为 10、20 和 30。

我们使用上述方法生成了训练集中所有实体对的 1－3 跳关系路径集合，并将这些关系路径集合和样本文本转换为矩阵形式，然后将这些矩阵输入 ConvF 进行模型训练，训练参数为：batch_size＝50；dropout_rate＝0.5；learning rate＝0.01；全连接隐藏层神经元个数＝200；关系向量的维度数＝20；关系抽取子模型设置高度为 2、3、4、5 的 4 个卷积核；针对 2 跳和 3 跳关系路径的卷积核数量＝200，特别地，对于单跳关系路径，由于我们仅考虑待预测关系 $r_{h\to t}$ 的反向路径 $r_{t\to h}$，这一反向路径只可能有|R|(即知识图谱中的关系种类数)种变化形式，故我们将单跳关系的卷积核数量设置为|R|。我们将 1－3 跳关系矩阵的高度分别设置为 5、20、60，这使得各矩阵可分别容纳最多 5、10、20 条关系路径的信息。另外，我们在模型训练过程中使用了早停法(early stopping)[28] 以减少过拟合的发生，即在训练过程中由系统根据验证损失(validation loss)是否上升，自动确定何时停止迭代。

3.4　结果与分析

鉴于当前相关研究主要是将外部文本特征与基于知识表示学习的关系预测方法相结合，我们选择基于知识表示学习方法中常用的两个翻译模型(TransE 和 TransH)和两个语义匹配模型(DistMult 和 ComplEx)作

为对照模型,与我们提出的ConvF模型进行关系预测的性能对比。表2和表3展示了ConvF模型和上述4个对照模型分别在FB15K-T和FB15K237-T评测数据集上的关系预测评测结果。其中,对照模型的性能由我们利用OpenKE[29]工具包进行关系预测后统计得到。由于4个对照模型都是基于知识图谱的三元组结构进行关系预测,故在评测时仅使用FB15K-T和FB15K237-T数据集中四元组的前三元组信息进行训练与预测。ConvF模型的性能则使用本文提出的融合内部结构和外部文本信息的方法,利用这两个数据集中的完整四元组信息进行关系预测统计得到。表2和表3的数据显示,ConvF模型在两个数据集的各评测指标上均明显超过对照模型。

表2　FB15K-T数据集的关系预测性能对比

模型	MRR		Hits@1		Hits@3		Hits@5	
	raw	filter	raw	filter	raw	filter	raw	filter
TransE	0.720	0.870	0.565	0.805	0.858	0.925	0.937	0.954
TransH	0.726	0.885	0.569	0.826	0.869	0.936	0.942	0.958
DistMult	0.462	0.504	0.244	0.285	0.615	0.676	0.736	0.782
ComplEx	0.633	0.740	0.483	0.645	0.750	0.816	0.832	0.862
ConvF	0.781	0.936	0.642	0.899	0.915	0.969	0.969	0.981

表3　FB15K237-T数据集的关系预测性能对比

模型	MRR		Hits@1		Hits@3		Hits@5	
	raw	filter	raw	filter	raw	filter	raw	filter
TransE	0.864	0.873	0.798	0.815	0.920	0.920	0.946	0.947
TransH	0.874	0.883	0.812	0.831	0.927	0.928	0.951	0.952
DistMult	0.525	0.528	0.411	0.415	0.585	0.588	0.658	0.660
ComplEx	0.505	0.509	0.401	0.408	0.552	0.554	0.628	0.629
ConvF	0.902	0.912	0.849	0.868	0.949	0.950	0.966	0.967

在上面的评测中,将1+2+3跳关系路径信息和文本信息同时作为模型输入。一个值得探讨的问题是单项特征对模型整体性能的贡献以及不同特征组合对模型性能的影响。为分析这一问题,我们尝试使用不同单项特征及其组合作为模型输入,然后在FB15K-T和FB15K237-T数据集上对ConvF模型的性能(基于MRR和Hits@1指标)重新进行评测,结果如表4所示(特征栏中1、2、3、T分别代表1-3跳路径特征和文本特征):

表4　不同特征和特征组合对ConvF性能的影响

特征	FB15K-T				FB15K237-T			
	MRR		Hits@1		MRR		Hits@1	
	raw	filter	raw	filter	raw	filter	raw	filter
only 1	0.649	0.749	0.528	0.691	0.203	0.203	0.087	0.087
only 2	0.521	0.595	0.387	0.503	0.592	0.598	0.512	0.524
only 3	0.673	0.766	0.528	0.671	0.854	0.864	0.790	0.808
only T	0.674	0.813	0.510	0.730	0.749	0.758	0.656	0.673
1+2	0.729	0.848	0.597	0.792	0.596	0.602	0.503	0.513
1+2+3	0.763	0.882	0.631	0.825	0.865	0.873	0.804	0.820
1+2+3+T	0.781	0.936	0.642	0.899	0.902	0.912	0.849	0.868

表4的结果显示,在单项特征中,3跳关系路径特征和文本特征相对具有较佳的表现。由于FB15K237/FB15K237-T数据集在FB15K/FB15K-T的基础上去除了实体间的直接反向关系,而ConvF在生成单跳关系路径时仅考虑实体对的直接反向关系,故仅使用单跳关系路径特征作为输入的ConvF模型在FB15K237-T数据集上的表现远低于FB15K-T数据集。尽管如此,ConvF仍然可以利用FB15K237-T中的2、3跳关系路径特征和文本特征取得不错的预测性能。将多种特征的组合作为模型输入,其性能总体优

于使用单一特征，且模型性能随着特征的叠加而递增。仅依靠 1 + 2 + 3 跳关系路径特征，ConvF 模型即可实现与 TransE、TransH 等主流模型相当的预测性能，而在叠加关系路径特征和文本特征的情况下，ConvF 模型的预测性能则超过了主流模型，这显示添加文本特征对于模型性能的提升具有较为明显的效果。

4 应用案例

本节以人物知识图谱的关系预测为例，说明 ConvF 模型的应用模式与应用价值。在百度百科人物词条的信息框中包含部分人物关系信息，如图 8 为“李隆基”词条中所展示的人物关系，这些人物关系信息对于知识服务具有重要价值。然而，这些已明确标注的人物关系仍远不能涵盖百度百科涉及的众多人物关系，其中存在着大量信息的缺失，例如，根据图 8 中的人物关系，易于估计“李宪”和“李治”间存在祖孙关系，而这一关系在两人物词条中并未被明确标注。



图 8 “李隆基”词条中标注的人物关系信息

EncycloGraph 是我们基于百度百科构建的一个大规模人物知识图谱，本研究利用 ConvF 模型补全该知识图谱中的人物关系。具体过程为：首先通过爬虫从百科人物词条页面抽取已明确标注的人物关系，共得到 43 682 个人物对间的关系；然后，利用此类已明确标注关系类型的人物对生成关系路径集合矩阵，利用百科中包含这些人物对指称的句子集合生成文本矩阵，将两类矩阵作为训练数据导入 ConvF 模型进行模型训练。在此前的研究中，我们已对百科人物的相关度进行了计算[30]，本研究利用训练出的关系预测模型对知识图谱中具有较高相关度，但并未被明确标注关系类型的人物对进行了关系预测，若模型判定人物对属于特定关系类别的概率超过一定阈值，则将该人物对标注为具有该人物关系。

通过上述方法，共预测出 106 770 对百科人物间的具体关系，从而使人物知识图谱具有丰富的人物关系类型信息。利用这一知识图谱，可以动态生成符合特定关系类型的人物关系网络(如亲缘网络、合作网络或学缘网络)，例如，图 9 显示了公元 420 年 – 公元 589 年(即南朝时期)重要人物的亲缘关系网络(即网络中的人物关系类型属于任意一种亲属关系)。可以看出，ConvF 模型可有效应用于知识图谱的补全，进而提升知识服务的质量。

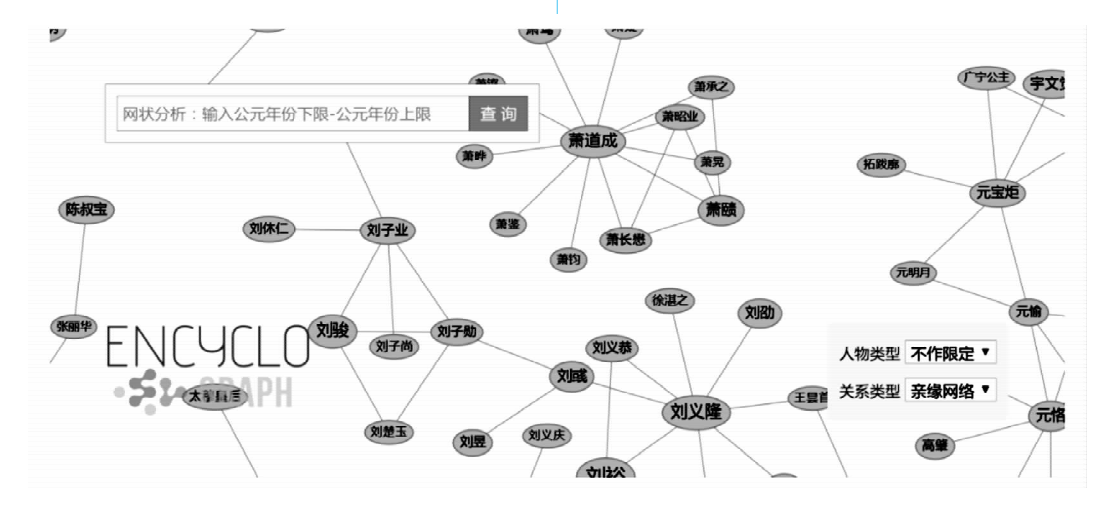


图 9 利用关系预测结果生成的亲缘关系网络示例

5 结语

本研究提出了一种融合结构与文本特征进行知识图谱关系预测的新方法。该方法的创新点在于：①设计了一种基于卷积神经网络的关系路径模式学习算法，该算法可利用深度学习技术实现更加有效的结构特征学习；②将知识图谱的关系路径结构特征与外部文本特征相结合，提出了一种特征融合的知识图谱关系预测模型 ConvF。

对比实验结果表明，ConvF 模型在评测数据集上

各评测指标的表现超过了对照主流模型的水平,这证明本研究所提出的融合模型具有良好的关系预测性能。我们测试了不同特征和特征组合对模型性能的影响,结果显示多特征叠加,特别是在结构特征的基础上叠加文本特征,可有效提升模型的预测性能。最后,本研究应用 ConvF 模型对人物知识图谱中部分缺失的人物关系进行了预测,预测效果显示 ConvF 模型在知识服务中具有良好的应用价值。该模型的不足之处在于其依赖于对知识图谱关系路径模式特征的学习,因此不适用于过于稀疏的知识图谱;另外,并非所有的知识图谱都易于找到与实体关系对应的文本信息,这使该模型的适用领域受到一定的限制,针对上述问题的改进还有待于未来的进一步研究。

参考文献:

[1] KEJRIWAL M. Domain-specific knowledge graph construction[M]. Berlin: Springer, 2019.

[2] 孙雨生,常凯月,朱礼军. 大规模知识图谱及其应用研究[J]. 情报理论与实践, 2018, 41(11): 138-143.

[3] KROMPAβ D, BAIER S, TRESP V. Type-constrained representation learning in knowledge graphs[C]// Proceedings of international Semantic Web conference. Berlin: Springer, 2015: 640-655.

[4] FAN M, ZHOU Q, ZHENG T F, et al. Distributed representation learning for knowledge graphs with entity descriptions[J]. Pattern recognition letters, 2017, 93(7): 31-37.

[5] 复旦大学知识工场实验室. 中文通用百科知识图谱(CN-DBpedia)[EB/OL]. [2020-07-21]. http://openkg.cn/dataset/cndbpedia.

[6] NGUYEN D Q, SIRTS K, QU L, et al. Neighborhood mixture model for knowledge base completion[C]// Proceedings of the 20th SIGNLL conference on computational natural language learning. Stroudsburg: ACL Press, 2016: 40-50.

[7] 徐增林,盛泳潘,贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4): 589-606.

[8] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261.

[9] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Proceedings of advances in neural information processing systems. San Diego: Neural Information Processing Systems Foundation, 2013: 2787-2795.

[10] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the 28th AAAI conference on artificial intelligence. Palo Alto: AAAI Press, 2014: 1112-1119.

[11] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the 28th AAAI conference on artificial intelligence. Palo Alto: AAAI Press, 2015: 2181-2187.

[12] JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]// Proceedings of the 53rd annual meeting of the ACL and the 7th international joint conference on NLP. Stroudsburg: ACL Press, 2015: 687-696.

[13] 聂斌玲. 基于图结构信息的知识表示学习方法研究[D]. 杭州:浙江大学, 2019.

[14] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]// Proceedings of the 28th international conference on machine learning. New York: ACM Press, 2011: 809-816.

[15] YANG B, YIH S W, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases[EB/OL]. [2020-08-16]. https://arxiv.org/pdf/1412.6575.

[16] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]// Proceedings of the 33rd international conference on machine learning. New York: ACM Press, 2016: 2071-2080.

[17] LAO N, MITCHELL T, COHEN W W. Random walk inference and learning in a large scale knowledge base[C]// Proceedings of the 2011 conference on empirical methods in natural language processing. Stroudsburg: ACL Press, 2011: 529-539.

[18] 刘峤,韩明皓,江浏祎,等. 基于双层随机游走的关系推理算法[J]. 计算机学报, 2017, 40(6): 1275-1290.

[19] XU J, CHEN K, QIU X, et al. Knowledge graph representation with jointly structural and textual encoding[C]// Proceedings of the 26th international joint conference on artificial intelligence. San Rafael: Morgan & Claypool Publishers, 2017: 1318-1324.

[20] LONG T, LOWE R, CHEUNG J C, et al. Leveraging lexical resources for learning entity embeddings in multi-relational data[EB/OL]. [2020-08-24]. https://arxiv.org/pdf/1605.05416.

[21] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2d knowledge graph embeddings[C]// Proceedings of the 32th AAAI conference on artificial intelligence. Palo Alto: AAAI Press, 2018: 1811-1818.

[22] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[EB/OL]. [2020-05-22]. https://arxiv.org/pdf/1207.0580.

[23] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Multi-instance multi-label learning for relation extraction[C]// Proceedings of the 2012 joint conference on EMNLP and computational natural language learning. Stroudsburg: ACL Press, 2012: 455-465.

[24] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 conference on empirical methods in natural language processing. Stroudsburg: ACL Press, 2014: 1746-1751.

[25] ADNAN WA, YAAKOB M, ANAS R, et al. Artificial neural network for software reliability assessment[C]// Proceedings of intelligent systems and technologies for the new millennium. Piscataway: IEEE, 2000: 446 - 451.
[26] TOUTANOVA K, CHEN D. Observed versus latent features for knowledge base and text inference[C]// Proceedings of the 3rd workshop on continuous vector space models and their compositionality. Stroudsburg: ACL Press, 2015: 57 - 66.
[27] Microsoft. FB15K - 237 knowledge base completion dataset[EB/OL]. [2020 - 08 - 10]. https://www.microsoft.com/en-us/download/details.aspx? id = 52312.
[28] BISONG E. Building machine learning and deep learning models on google cloud platform[M]. Ottawa: Apress, 2019.
[29] HAN X, CAO S, LV X, et al. Openke: an open toolkit for knowledge embedding[C]// Proceedings of the 2018 conference on empirical methods in natural language processing. Stroudsburg: ACL Press, 2018: 139 - 144.
[30] 林泽斐,欧石燕. 基于在线百科的大规模人物社会网络抽取与分析[J]. 中国图书馆学报, 2019, 45(6): 100 - 118.

作者贡献说明:

林泽斐:研究方案设计、数据处理和论文撰写;

欧石燕:论文审阅、修改和定稿。

Research on Relation Prediction in Knowledge Graphs by Fusing Structure and Text Features

Lin Zefei[1,2] Ou Shiyan[1]

[1] School of Information Management, Nanjing University, Nanjing 210093

[2] College of Social Development, Fujian Normal University, Fuzhou 350007

Abstract: [**Purpose/significance**] Relation prediction is an important task in knowledge graph completion, and plays an important role in improving the completeness of knowledge in knowledge graphs. The paper proposes a new relation prediction method that combines internal structure features and external text features, which aims to predict the missing relations between two entities in knowledge graphs. [**Method/process**] The method transforms the relation paths in a knowledge graph and the texts that involve entity relationships into matrixes, learns the structure features and text pattern features related to a specific relation type through convolutional neural networks, and then trains a model based on the learned features for relation prediction. [**Result/conclusion**] The results shows that the performance of our proposed method on evaluation data sets is superior to the state-of-the-art approaches, and the method can effectively improves the performance of knowledge graph relationship prediction. Through practical application, it wvas found that this method has high application value in knowledge services.

Keywords: knowledge graph relation prediction feature fusion deep learning